IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 11 November 2024, accepted 22 December 2024, date of publication 25 December 2024,
date of current version 31 December 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3522465

RESEARCH ARTICLE

MIRoBERTa: Mental Illness Text Classification With Transfer Learning on Subreddits

MAVIN SAO[1] **AND HOI-JEONG LIM**[1,2]
[1]Graduate School of Data Science, Chonnam National University, Gwangju 61186, Republic of Korea
[2]Public Data Analytics Center, Chonnam National University, Gwangju 61186, Republic of Korea

Corresponding author: Hoi-Jeong Lim (hjlim@jnu.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) through Korean Government (MSIT) 50% under Grant RS-2023-00242528; and in part by the Innovative Human Resource Development for the Local Intellectualization Program through the Institute of Information and Communications Technology Planning and Evaluation (IITP), Korean Government (MSIT) 50% under Grant IITP-2024-RS-2022-00156287.

ABSTRACT Social media has emerged as a critical resource for text classification, with Reddit prominent among these platforms. In addition to serving as a space for users to share thoughts openly, Reddit is a substantial repository for storing diverse expressions, encompassing discourse on mental health. This study employs the Reddit PullPush application programming interface to gather posts related to seven common mental illnesses: attention-deficit/hyperactivity disorder, anxiety, bipolar disorder, borderline personality disorder, depression, obsessive-compulsive disorder, and post-traumatic stress disorder. A robust collection of 54 161 submissions across 11 subreddits was collected, with the ''none'' class indicating the absence of mental illness. This study comparatively evaluated three traditional machine learning, two powerful bidirectional deep learning models, and four transformer models: bidirectional encoder representations from transformers (BERT), robustly optimized BERT pretraining approach (RoBERTa), BigBIRD, and long-document transformer (Longformer) on a multiclass text classification task. The BigBIRD and Longformer approaches outperformed standard transformers, BERT and RoBERTa, on the multiclass text classification task. BigBIRD achieved the highest accuracy and F1-score of 0.840, whereas the RoBERTa model had a slightly lower accuracy and F1-score of 0.834. We pretrain the RoBERTa-base and BERT-base-uncased model using a fill-mask task in the public mental illness corpus domain to improve language understanding before fine-tuning. This pretrained model, mental illness RoBERTa (MIRoBERTa), outperformed other models on the text classification task with an accuracy of 0.847 and F1-score of 0.847. Additionally, mental illness BERT (MIBERT) surpassed existing domain-specific pretrained models with an accuracy of 0.835 and an F1 score of 0.835. We also explore the effectiveness of ensemble techniques by combining the domain adaptation model with the original variant. Finally, we analyze word importance to identify the terms that most significantly contribute to the model classification decisions.

INDEX TERMS Mental health, natural language processing, text classification, transfer learning, transformers.

I. INTRODUCTION

Since 2019, the infectiousness of the coronavirus disease 2019 (COVID-19) has posed challenges to global health systems regarding in-person operations [1]. As remote systems for monitoring physical health have increased, the demand for similar systems for mental health management has also increased. In addition to the psychological side effects of COVID-19, one in eight patients has expressed concerns regarding the absence of an effective vaccine, and the adverse socioeconomic repercussions (e.g., unemployment) have emphasized the necessity of addressing mental health problems [2], [3]. In response to this urgent need, remote systems have been implemented to cater to mental health needs.

The associate editor coordinating the review of this manuscript and approving it for publication was Arianna Dulizia.

Many studies have indicated that social media platforms (e.g., Facebook, X, and Reddit) are often employed by individuals as channels to express their mental health [4], [5]. These platforms are valuable applications for sharing opinions and venting. According to a 2023 study [6], social media platforms have been recognized as potentially beneficial tools in mental health care. They are repositories of information, facilitating access to valuable resources. Participating in online groups or communities can offer a sense of belonging, emotional support, and the opportunity to share experiences with people facing similar challenges.

Reddit has become a leading resource for text mining among researchers. The dataset in this study was constructed using Reddit with the PullPush application programming interface (API) [7] to collect data from subreddits that focus on mental health disorders. A typical approach is text-mining tasks using text as a resource in natural language processing (NLP) to address escalating mental health challenges. Advanced pretraining language models can mine text and extract information.

In this study, the main aim is to improve the contextual understanding of mental health-related text through domain-adaptive pretraining of language models on mental health-specific data from Reddit. This study makes the following contributions:

- Domain-specific masked language models (MIBERT and MIRoBERTa) are developed based on BERT and RoBERTa architectures. While these base models are powerful, they are limited by their general-purpose pretraining, which affects their ability to accurately predict mental health-specific and negative sentiment words in masked language modeling tasks.
- A comprehensive fine-tuning approach is implemented using a mental illness corpus, which enhances the models' capabilities to predict and understand domain-specific vocabulary and emotional context particularly in mental health discussions.
- To address the challenge of processing long text sequences common in mental health discussions, the study incorporates and analyzes the performance of advanced architectures like BigBird and Longformer.
- As demonstrated in our experimental section, our modified models outperform not only their original architectures (BERT and RoBERTa) but also other domain-adaptive pretrained models across various mental illness classification tasks.
- The study provided a detailed analysis of influential words that impact model decisions, helping identify specific words that strongly influence predictions and enhance our understanding of language patterns associated with mental health classifications.
- To demonstrate practical applications and ensure accessibility, we've deployed our model on Hugging Face Spaces and developed a Streamlit-based web application for mental illness prediction from text. Additionally, the dataset and trained models are made publicly available to support future research in this domain.

The remainder of this paper is organized as follows. Section II reviews related research. Section III introduces our dataset. Section IV outlines our experimental design. Section V describes the experimental setup. Section VI presents our experimental results and discusses them. Finally, Section VII concludes the study.

II. RELATED WORK

Recent studies have applied machine and deep learning techniques to classify mental health-related text. A study [8] investigated the problem of detecting suicidal ideation on social media platforms (i.e., Reddit and X) by applying feature processing and traditional supervised classification. Two neural network models were used in addition to four traditional supervised learning methods. The extreme gradient boosting method outperformed other models in accuracy, recall, F1-score, and area under the receiver operating characteristic curve, followed by long short-term memory (LSTM), which outperformed all models in precision. In 2018 [9], researchers assessed a task to detect users with signs of depression by employing deep learning techniques to classify it in X texts. They proposed a method of determining an efficient neural network architecture to improve and optimize word embeddings. Moreover, the authors compared models based on convolutional neural networks and recurrent neural networks to determine the best model to detect depression. In 2019 [10], NLP techniques and machine learning approaches were employed to examine Reddit user posts to detect factors that could expose the attitudes of relevant online users toward depression. Another study [11] employed multinomial naive Bayes and support vector regression algorithms as classifiers to analyze health tweets regarding depression and anxiety from mixed tweets. Moreover, the work in [12] employed an ensemble method that applies deep learning techniques to classify text from Reddit posts with signs of anxiety and depression. In [13], the author applied two models: the convolutional neural network with bidirectional LSTM (BiLSTM) and extreme gradient boosting. The study aimed to identify and understand suicidal ideation patterns in social media posts. In 2020 [14], NLP techniques were employed to analyze content on Reddit. The aim was to explore how discussions on Reddit changed during the beginning of the COVID-19 pandemic by focusing on 15 major mental health support groups and comparing them to 11 non-mental health groups.

Recent advances in time series forecasting introduced the GinAR [15] Network, which handles missing data using interpolation attention and adaptive graph convolution. In distributed deep learning [16], adaptive communication compression reduces training overhead by adjusting data compression dynamically. Both innovations enhance model robustness and scalability for incomplete data and large-scale models.

The text classification field has rapidly evolved, shifting from traditional machine learning techniques. The bidirectional encoder representations from transformers (BERT), a groundbreaking language model released in 2018, transformed the NLP field by performing better on diverse tasks, such as general language understanding evaluation and multigene natural language inference accuracy [17]. Its success is primarily attributed to pretraining on a massive collection of textual data, including books and Wikipedia. In [18], the robustly optimized BERT pretraining approach (RoBERTa) was applied to detect and classify signs of five mental illnesses. Moreover, BERT-like models are adaptable to specialized domains, such as biomedicine. For example, biomedical BERT (BioBERT), pretrained on biomedical literature, demonstrates superior performance in biomedical text-mining tasks compared to the original BERT [19]. The success of BioBERT underscores the importance of domain-specific pretraining to achieve optimal results.

In another study [20], the depression classification task was approached by comparing four small transformer models: the Efficiently Learning an Encoder that Classifies Token Replacements Accurately (ELECTRA) small generator, ELECTRA small discriminator, XtremeDistil-L6, and a lite BERT (ALBERT) base V2 to classify depression intensity using tweets. The models were employed to classify three label classes: severe, moderate, and mild. The ELECTRA small generator outperformed other models and was proposed to be suitable for low-powered devices. In [21], RoBERTa language models were applied to classify Reddit posts into none, moderate, and severe depression. This approach, which included fine-tuning and developing a domain-specific model called depression RoBERTa (DepRoBERTa), successfully identified signs of depression.

However, many studies have focused only on specific mental illnesses, such as suicidal ideation, depression, and others. This study explores eight-class classification using the dataset we constructed. Moreover, adapting domain knowledge to a pretrained language model dramatically enhances the accuracy of multiclass classifications in mental health studies.

III. DATASET

A. COLLECTION METHOD

The dataset in this study was collected from Reddit using the PullPush API [7]. The approach involved collecting subreddit submissions related to mental health, including r/ADHD, r/anxiety, r/bipolar, r/BPD, r/depression, r/OCD, and r/PTSD. Posts with over ten upvotes were filtered when making API requests to ensure that at least ten people agreed on its relevance or importance in each subreddit. The same process was applied to collect submissions from r/Jokes, r/love, r/productivity, and r/happy unrelated to mental illness. Completing the collection resulted in a dataset of 76 584 submissions from these subreddits.

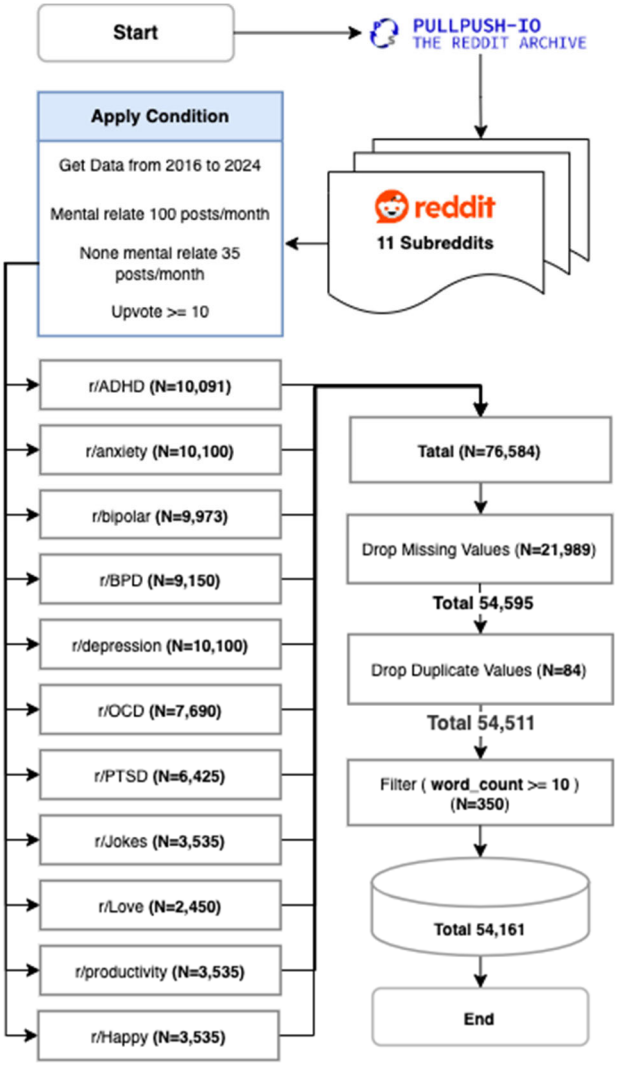

FIGURE 1. Flowchart of the data preparation process.

B. DATA CLEANING AND QUALITY ASSURANCE

We cleaned the data to remove duplicates and address missing values to ensure the quality and consistency of the dataset. All data instances were restricted to those containing a description with a word count of at least ten words to ensure that the user submissions contained substantial content. We obtained 45 954 submissions related to mental illness and 8207 unrelated to mental illness. Fig. 1 presents the processing details.

Finally, the data were combined and labeled based on subreddits for mental illness submissions, and entries unrelated to mental illness were labeled ''none''. This labeling approach resulted in a dataset of 5 161 submissions. The data were collected between January 2016 and July 2024, and any URLs, identifications, or usernames that contained potentially sensitive information were removed. Tables 1 and 2 present the dataset examples and descriptions, respectively.

IV. EXPERIMENTAL DESIGN

We structured the solution into six steps, from data processing to ensemble learning, integrating several machine and deep

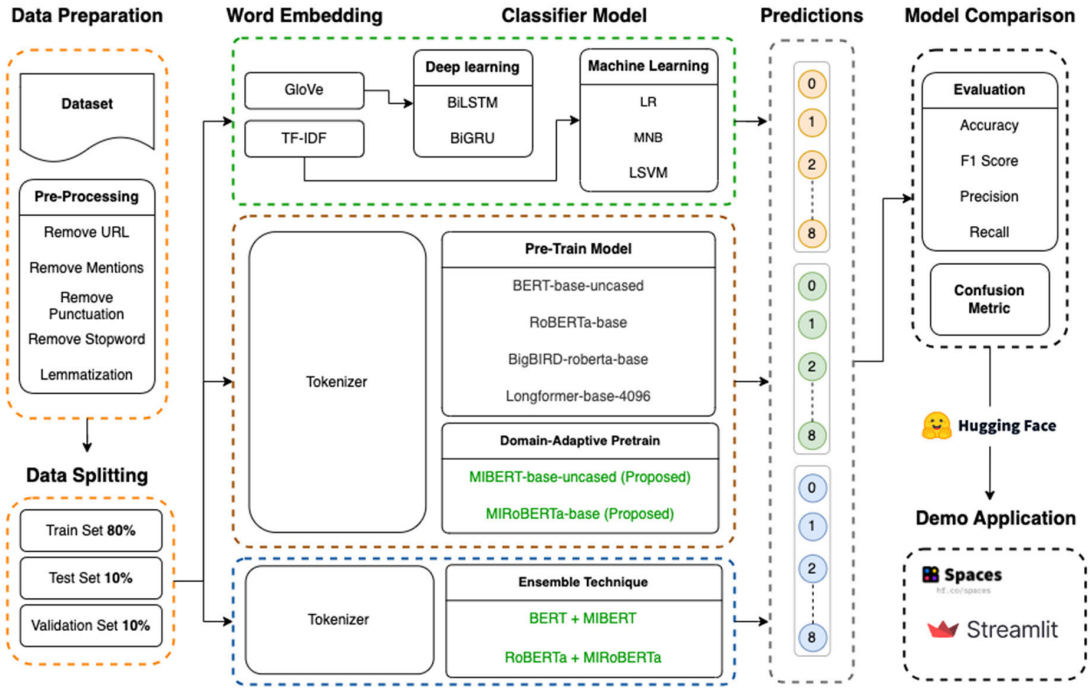

FIGURE 2. Overall architecture of the mental illness text classification system.

learning approaches to ensure significant results in this experiment and develop the baseline models. Each step is clarified below.

A. DATA PROCESSING

Effective data preprocessing is crucial in optimizing the proposed model via fine-tuning. The text was converted to lowercase, and URLs, punctuation, and user mentions were removed. This process improved the overall model by focusing on more meaningful words. We applied the lemmatization technique to reduce words to their base or dictionary forms. We did not remove stop words and applied the lemmatization technique to transformer models to help maintain the contextual integrity of text sequences [22].

B. TRADITIONAL MACHINE LEARNING MODEL

We used the term frequency-inverse document frequency (TF-IDF) [23] to create numerical representations of the words in the documents. The TF-IDF method helps identify the most critical words in each document while reducing the influence of common words frequently appearing across all documents. We chose three well-known machine learning models for the classification task: logistic regression, multinomial naive Bayes, and linear support vector machine (LSVM). These models were selected because they work well as starting points in machine learning projects.

- **Logistic regression** is simple and easy to interpret.
- **Multinomial naive Bayes** works well for classifying text because it applies probability.
- The **LSVM** handles high-dimensional data (with many features) well.

We separately trained each model to assess how well the model could correctly categorize mental health classifications. By doing this, we established a solid baseline for comparing the performance of various approaches.

C. DEEP LEARNING MODEL

For deep learning, we employed the BiLSTM and bidirectional gated recurrent unit (BiGRU) models due to the effectiveness of the bidirectional architecture in text classification. In this experiment, we employed global vectors (GloVe) for word representation embeddings [24] to enhance the ability of the models to understand and represent semantic relationships in the text, enriching the input representations.

1) GLOBAL VECTORS FOR WORD REPRESENTATION

The global vectors for word representation (GloVe) model converts words into high-dimensional vectors by applying co-occurrence statistics from extensive text corpora. These embeddings provide a nuanced understanding of word meanings and relationships, which is crucial for efficient text classification. By integrating GloVe embeddings in the deep learning model, we use pretrained vectors that encapsulate rich linguistic information. This approach ensures that, even with smaller datasets, the models can generalize and perform better because the embeddings carry substantial knowledge regarding the language.

2) BIDIRECTIONAL LONG SHORT-TERM MEMORY

A BiLSTM is a recurrent neural network that processes text sequences forward and backward, capturing context from both directions. This bidirectional processing enables the

TABLE 1. Text example with label from dataset.

Text	Label
Sometimes, I get so bored that it feels like nothing in the world could ever excite me: I'll ask myself, "If I could do anything right now, what would it be?" and nothing comes to mind. It's frustrating dealing with ADHD and boredom feels like a constant struggle. You want to feel alive, but you just don't know how.	ADHD
I haven't felt this good in a long time. Hi everyone! Today was almost miraculous: I didn't have a single anxiety attack, and I'm actually anxiety breathing like a normal person instead of like Darth Vader. I'm just so happy, and I want to say thanks for all the advice I've learned here.	Anxiety
Does anyone else get angry that other people exist?: I sometimes get into moods where I hate going out in public just because other people are there. It can actually put me in a bad mood. If I walk into a crowded store, I get especially annoyed and find myself wishing it was just me and my fp in the world.	BPD
I have OCD, so I like things neat: This might sound silly, but maybe you all can relate and laugh a little—I'm at work, and someone just complained about their desk being messy. Meanwhile, here I am, with the world's messiest desk, and it's all I can think about!	OCD
Friendly advice needed: I was recently diagnosed with bipolar disorder and was wondering if anyone else experiences anger and outbursts as a symptom. It's putting a strain on my relationship because every little thing that frustrates me can blow up into intense anger. Any advice on how to handle this would be really appreciated.	Bipolar
I would be dead if it weren't for my family: Literally the only thing keeping me alive is the fact that my mom would be completely broken if I died. It's what keeps me from ending it. I'm already a disappointment; I don't want to be the cause of more depression after I'm gone.	Depression
Reminders of the past: Does anyone else experience tiny triggers that might seem insignificant but end up keeping you up at night, making you remember things you've been working so hard to forget? I'll be sitting somewhere, see something, and suddenly I'm flooded with memories I'd rather leave behind.	PTSD
I got accepted to college: It's the first step toward my dream of becoming a math teacher! After spending the last few years serving my country, I'm incredibly excited for this change of pace. I'm going to throw my heart into this degree, and I'm so proud of myself for getting here.	None

TABLE 2. Dataset description.

Subreddit	No. of posts	%	Avg. word count
ADHD	8931	16.41%	194.94
Anxiety	8063	14.86%	162.28
Bipolar	4062	7.50%	135.76
BPD	7244	13.31%	173.83
Depression	7398	13.62%	158.76
OCD	4764	8.81%	147.76
PTSD	5492	10.10%	197.88
None	8207	15.38%	162.20

model to consider the surrounding words to understand the context more deeply, leading to more accurate classification. By addressing problems, such as vanishing gradients, BiLSTM can address long-term dependencies in the text.

3) BIDIRECTIONAL GATED RECURRENT UNIT

A bidirectional gated recurrent unit (BiGRU) is a sequence processing model designed to process text sequences in the forward and backward directions. This model uses special mechanisms called ''gates'' to control the information flow through the network. These gates facilitate remembering crucial details and forgetting irrelevant ones as the model processes the text. This approach makes BiGRU an excellent choice for text classification tasks where performance and computational efficiency are critical. By capturing the comprehensive context, BiGRUs provide speed and accuracy in classification.

D. TRANSFORMER: DOWNSTREAM TASK FINE-TUNING MODEL

Four commonly used pretrained language transformer models were fine-tuned: BERT [17], RoBERTa [25], BigBIRD [26], and Longformer [27]. Fine-tuning these models on downstream tasks is essential to adapting them to specific domains and improving their performance on targeted applications. Each model was selected based on the specific needs of the text classification task and the dataset characteristics.

1) BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMER

The BERT model is a revolutionary breakthrough in NLP, consistently achieving top performance on diverse tasks. In this implementation, the BERT-base-cased model was employed for tokenization and classification. Its bidirectional training approach considers context from left-to-right and right-to-left, allowing for a nuanced understanding of complex language structures and making it an effective baseline for evaluating other models. Additionally, the lower demand of BERT for computational resources than newer models makes it practical when balancing performance with resource availability.

2) ROBUSTLY OPTIMIZED BERT PRETRAINING APPROACH

The RoBERTa model is a leading language model in the field of NLP, which builds on the groundbreaking BERT architecture, introducing refinements to crucial hyperparameters and extensive training on a massive dataset.

The model focuses on masked language modeling (MLM), removing next-sentence prediction tasks and enhancing its performance for deep contextual understanding. With access

TABLE 3. Hyperparameters for domain-adaptive pretraining model.

Model	Num train epochs	Per device train batch size	MLM probability	Block size
BERT-base-uncased	10	16	0.15	256
RoBERTa-base	10	16	0.15	256

to significant computational resources, training RoBERTa on a more extensive and diverse dataset allows for superior generalization and higher accuracy in the text classification task. This study uses the pretrained RoBERTa-base model as the basis for the tokenization process and core classification model.

3) BIGBIRD: TRANSFORMERS FOR LONGER SEQUENCES

The BigBIRD model enhances the ability of the transformer to process longer sequences by incorporating sparse and global attention. The combination of sparse and global attention mechanisms allows BigBIRD to capture long-range dependencies while remaining computationally efficient. Including BigBIRD in this study enables comparing its performance with established models, such as BERT and RoBERTa, particularly in handling longer sequences. The BigBIRD design supports scalable processing, making it well-suited for this dataset, which includes documents of varying lengths.

4) LONGFORMER: LONG-DOCUMENT TRANSFORME

Longformer is designed to handle long text sequences and can process large documents using sparse attention mechanisms without the typical computational challenges associated with longer sequences. This approach makes it especially suitable for datasets containing longer text sequences. Unlike traditional transformers, such as BERT and RoBERTa, which perform poorly with longer documents due to their more complex attention processes, Longformer works well with extended texts.

E. DOMAIN-ADAPTIVE PRETRAINING

We employed a Reddit mental health dataset to build the corpus of textual data relevant to mental illness [28], [14]. The dataset of 243 335 submissions contains extensive data valuable for language model pretraining. These data were obtained with the following labels: depression (46.7%), anxiety (22.1%), ADHD (17%), borderline personality disorder (BPD) (9%), PTSD (3.1%), and bipolar (2.1%).

Instead of training the language model from scratch, we adapted transfer learning using the checkpoints from the original RoBERTa and BERT models. The pretrained language transformer models, RoBERTa-base and BERT-base-uncased, were employed to retrain the constructed corpus for better performance on a downstream task. The training was conducted using an Nvidia H100 Tensor Core graphics processing unit with a batch size of ten per device and 259 500 steps. The training for RoBERTa takes about 9 h to complete; for BERT, it takes about 8 h and 30 min. Table 3 presents the pretraining hyperparameter details. This pretraining approach optimizes the ability of the mental illness RoBERTa (MIRoBERTa) and mental illness BERT (MIBERT) to understand text in the mental health domain.

1) MASKED LANGUAGE MODELING AND GRADIENT-BASED PARAMETER UPDATE

A critical technique used during pretraining is MLM, where random parts of the input text are hidden, and the model must predict missing words. The loss function (1) measures the accuracy of the model prediction of masked tokens, guiding the pretraining process

$$\mathrm{L_{MLM}} = -\frac{1}{M}\sum_{i=1}^{M} \log \mathrm{P}\left(w_i = w_i^* \mid w_{\setminus i}\right) \tag{1}$$

where M represents the total masked tokens in a single training sequence, and w_i signifies the original, true token at position i in the sequence. Conversely, $w*_i$ denotes the predicted token at the same position i. Finally, $w_{\setminus i}$ catches the entire context of the sentence, excluding the masked token at position i. This context is critical for the model to make accurate predictions regarding the masked words. The model parameters are updated using gradient-based optimization (2) to improve the ability of the domain-adaptive pretraining model to understand the context of mental health language:

$$\theta_t = \theta_{t-1} - \alpha \cdot \nabla_\theta \mathrm{L_{MLM}} \tag{2}$$

where θ_t represents the model parameters at the current training stage (time step t), and θ_{t-1} denotes the model parameters at the previous time step. The difference between these two parameter sets reflects the training adjustments. The learning rate α controls the magnitude of these adjustments, acting as a scaling factor to determine the extent to which the model should adapt its parameters in response to the calculated gradient of the MLM loss, $\nabla_\theta \mathrm{L_{MLM}}$. This gradient improves performance by indicating the direction and magnitude of the parameter change.

2) MASKED LANGUAGE MODELING LOSS MINIMIZATION ON DOMAIN-ADAPTIVE PRETRAINING PERFORMANC

Minimizing this loss contributes to the overall improvement in understanding mental health language using MIBERT and MIRoBERTa in several critical ways:

- **Contextual Understanding**: Minimizing the MLM loss facilitates a better understanding of the model in the context in which mental health-related terms are used. This understanding is crucial for accurately interpreting

subtle language (e.g., word choice, context) often found in mental health texts.

- **Semantic Precision**: By learning to predict masked tokens accurately, the model gains a deeper understanding of the semantics of words and phrases specific to mental health, improving its ability to identify subtle differences in meaning.
- **Enhanced Generalization**: As the model enhances its ability to anticipate masked tokens, it cultivates a more resilient understanding of the language, allowing it to generalize more effectively to various downstream tasks, including sentiment analysis and symptom identification, which are crucial in mental health.
- **Domain-Specific Adaptation**: The specialized training on mental health texts helps the model learn the unique language and terms in this field. This learning makes predictions more accurate and reliable for mental health applications.

These improvements enhance the performance of MIBERT and MIRoBERTa in understanding, classifying, and interpreting mental health-related content, making it a valuable tool for mental health applications.

F. ENSEMBLE LEARNING

Ensemble learning is a machine learning approach combining the predictions of multiple models to improve the overall performance. Ensemble learning techniques were applied by grouping models into two ensembles: one consisting of RoBERTa and MIRoBERTa and the other consisting of BERT and MIBERT using the ensemble average [29]. Ensemble averaging for each group is defined as follows:

$$\text{Ensemble Ave. RoBERTa} = \frac{M_{RoBERTa} + M_{MiRoBERTa}}{2} \quad (3)$$

$$\text{Ensemble Aver. BERT} = \frac{M_{BERT} + M_{MIBERT}}{2} \quad (4)$$

where M_{RoBERTa}, $M_{\text{MIRoBERTa}}$, M_{BERT}, and M_{MIBERT} represent the predictions from the respective models in the subscripts. Fig. 2 illustrates the overall process.

V. EXPERIMENTAL SETUP

This section details each model configuration for the multiclass mental illness classification task. The experiment employs a diverse set of models for a comprehensive comparison. Machine learning models were established as a traditional baseline approach. Including BiLSTM and BiGRU provides insight into the performance of various recurrent neural network architectures. All models were implemented in PyTorch [30] and the Hugging Face Transformers Library [31].

A. DEEP LEARNING MODEL CONFIGURATION

We developed multilayer BiLSTM and BiGRU networks in this implementation. Both architectures have a 300-dimensional embedding layer using pretrained GloVe embeddings (glove.6B.300d.txt), with a maximum sequence length of 1024 tokens. The BiLSTM and BiGRU configurations are identical except for the recurrent layer type. Each network comprises two bidirectional recurrent layers (either LSTM or GRU) with 256 and 128 units, respectively, followed by dropout layers with a rate of 0.25 for regularization [32]. A dense layer, with 64 units and rectified linear unit activation, precedes the final output layer, which employs softmax activation for multiclass classification. The models apply the Adam optimizer with a configurable learning rate and are compiled using a sparse categorical cross-entropy loss function.

B. HYPERPARAMETER TUNING

All models were trained on a combination of title and post-text data via the cross-entropy loss function. The BiLSTM and BiGRU models were trained for five epochs with a batch size of 64 and testing learning rates of 1e-3, 2e-3, and 3e-3. The AdamW optimizer was used for transformer models, with training varying based on their architecture. The BERT and RoBERTa models used a maximum sequence length of 512 tokens, whereas BigBIRD and Longformer processed longer sequences of up to 1024 tokens. All transformer models were trained for three epochs with a batch size of 16, with learning rates of 1e-5, 2e-5, and 3e-5. This setup allows a comprehensive comparison across model types and hyperparameters.

C. EVALUATION METRICS

We assessed performance using classification metrics (accuracy, precision, recall, and the F1-score), ensuring a reliable model evaluation. These metrics rely on a few critical concepts:

- **True positive (TP):** The model correctly predicts a text as belonging to a specific mental health class.
- **True negative (TN):** The model correctly predicts a text as not belonging to a specific mental health class.
- **False positive (FP):** The model incorrectly predicts a text as belonging to a specific mental health class (when it belongs to a different class).
- **False negative (FN):** The model incorrectly predicts a text as not belonging to a specific mental health class (when it does belong).

These values were extracted from the confusion matrix and were calculated as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (5)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (6)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (7)$$

$$\text{F1Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (8)$$

The reported results are presented as weighted-average scores to account for class imbalances, ensuring a balanced

TABLE 4. Performance comparison of machine learning, transformer, and deep learning models by learning rate (LR).

Model	LR	Acc.	Pre.	Re.	F1
Machine learning model					
TF-IDF+Logistic Regression		0.771	0.777	0.771	0.770
TF-IDF + MNB		0.663	0.701	0.663	0.661
TF-IDF + LSVM		**0.772**	**0.776**	**0.772**	**0.772**
Deep learning model					
GloVe + Bi-LSTM	1e-3	0.772	0.783	0.772	0.773
	2e-3	**0.782**	**0.787**	**0.782**	**0.783**
	3e-3	0.775	0.780	0.775	0.775
GloVe + Bi-GRU	**1e-3**	**0.787**	**0.792**	**0.787**	**0.786**
	2e-3	0.775	0.782	0.775	0.775
	3e-3	0.772	0.786	0.772	774
Transformer model					
BERT	**1e-3**	**0.822**	**0.827**	**0.822**	**0.823**
	2e-3	0.817	0.821	0.817	0.817
	3e-3	0.812	0.818	0.812	0.812
RoBERTa	**1e-5**	**0.834**	**0.837**	**0.834**	**0.834**
	2e-5	0.826	0.834	0.826	0.827
	3e-5	0.823	0.828	0.823	0.823
BigBIRD	**1e-5**	**0.840**	**0.843**	**0.840**	**0.840**
	2e-5	0.837	0.839	0.837	0.836
	3e-5	0.835	0.827	0.835	0.835
Longformer	**1e-5**	**0.835**	**0.839**	**0.835**	**0.836**
	2e-5	0.817	0.831	0.817	0.818
	3e-5	0.824	0.827	0.824	0.824
Domain-adaptive pretrained model					
MIBERT (proposed)	**1e-5**	**0.836**	**0.839**	**0.836**	**0.836**
	2e-5	0.835	0.838	0.835	0.835
	3e-5	0.827	0.827	0.827	0.826
MIRoBERTa (proposed)	**1e-5**	**0.847**	**0.851**	**0.847**	**0.847**
	2e-5	0.841	0.843	0.841	0.840
	3e-5	0.837	0.840	0.837	0.837
Ensemble model					
Ensemble-BERT (proposed)		**0.845**	**0.848**	**0.845**	**0.845**
Ensemble-RoBERTa (proposed)		**0.851**	**0.855**	**0.851**	**0.851**

representation of the contribution of each class to the overall performance. This approach more accurately and fairly assesses model abilities across classes.

VI. EXPERIMENTAL RESULTS AND DISCUSSION

This section presents the evaluation results for all models, highlighting the best performance for each, discussing the error analysis, and identifying potential error sources. Finally, this section describes the deployment of the ensemble model in a simple web application for public testing.

A. EVALUATION RESULTS

Table 4 details the performance of the models on the mental illness classification challenge. Among the machine learning models, TF-IDF with LSVM performed best overall, with the highest accuracy (0.772) and F1-score (0.772). The TF- IDF method with logistic regression exhibited the best precision (0.777) among these models.

In the deep learning category, GloVe-based BiGRU slightly outperformed BiLSTM. Moreover, BiGRU achieved the best results with a learning rate of 1e-3, reaching an accuracy of 0.787 and an F1-score of 0.786. The BiLSTM model performed optimally at a 2e-3 learning rate, with an accuracy of 0.782 and an F1-score of 0.783. Among the transformer models, those designed for longer sequences displayed superior performance. In addition, BigBIRD attained the highest scores among standard transformers, with both the accuracy and F1-score at 0.840 at a 1e-5 learning rate. Longformer followed closely with an accuracy of 0.835 and an F1-score of 0.836 at the same learning rate. The BERT and RoBERTa models also performed well, with RoBERTa achieving slightly better results (accuracy and F1-score of 0.834) than BERT (0.823) at a 1e-5 learning rate.

The proposed MIRoBERTa models provided impressive results, with MIRoBERTa outperforming all individual models. The MIRoBERTa model achieved the highest individual model scores with both the accuracy and F1-score at 0.847 at a 1e-5 learning rate. The ensemble approaches further improved the results. The ensemble-RoBERTa model, combining RoBERTa and MIRoBERTa, achieved the best overall performance regarding the accuracy (0.851), precision (0.855), recall (0.851), and F1-score (0.851). This ensemble performance is a sizable improvement over individual models and establishes the effectiveness of the ensemble technique in this classification task.

Table 5 compares the proposed MIBERT model with other domain-specific BERT variants. The results demonstrate the effectiveness of the domain-adaptive pretraining approach. The MIBERT model outperforms other specialized models across all metrics, achieving the highest accuracy (0.835),

TABLE 5. Comparison of domain-adaptive pretrained models.

Model	LR	Acc.	Pre.	Re.	F1
BioBERT [19]	**1e-5**	**0.821**	**0.826**	**0.821**	**0.821**
	2e-5	0.808	0.808	0.804	0.807
	3e-5	0.806	0.811	0.806	0.806
ClinicalBERT [35]	1e-5	0.779	0.783	0.779	0.778
	2e-5	0.798	0.799	0.798	0.798
	3e-5	**0.799**	**0.803**	**0.799**	**0.799**
MedBERT [34]	1e-5	0.796	0.800	0.796	0.797
	2e-5	0.801	0.804	0.801	0.801
	3e-5	**0.804**	**0.806**	**0.804**	**0.804**
MentalBERT [33]	1e-5	0.831	0.835	0.831	0.831
	2e-5	**0.832**	**0.835**	**0.832**	**0.831**
	3e-5	0.819	0.823	0.819	0.816
MIBERT (Proposed)	**1e-5**	**0.835**	**0.837**	**0.835**	**0.835**
	2e-5	0.827	0.834	0.827	0.827
	3e-5	0.819	0.825	0.819	0.819

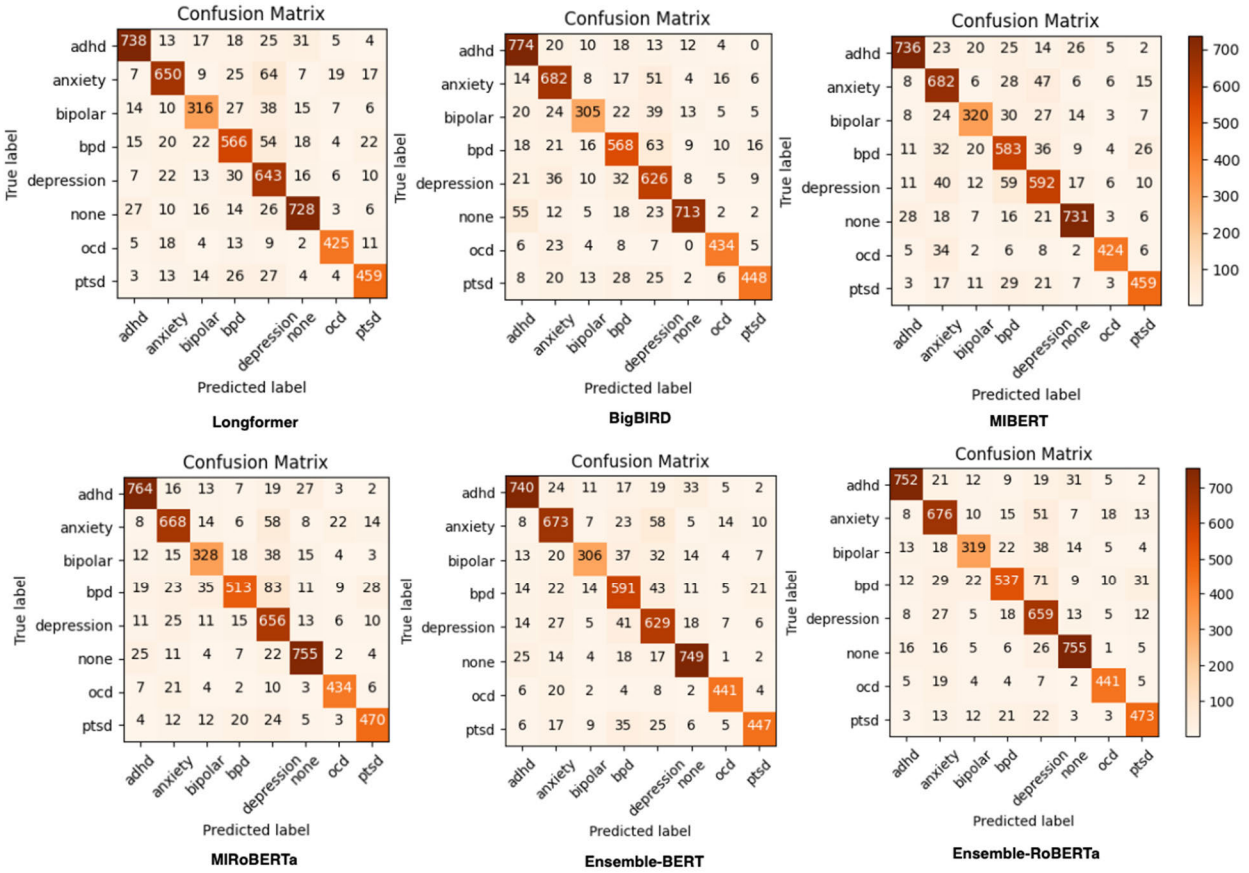

FIGURE 3. Confusion matrices: Domain-adaptive pretrained model vs. long-sequence transformer vs. ensemble models.

TABLE 6. Classwise results: Domain-adaptive pretrained model vs. long-sequence transformer vs. ensemble models.

Class	Longformer			BigBIRD			MIBERT		
	P	R	F1	P	R	F1	P	R	F1
ADHD	0.904	0.867	0.885	0.845	**0.910**	0.876	0.909	0.865	0.886
Anxiety	**0.860**	0.815	0.837	0.814	**0.855**	0.834	0.784	**0.855**	0.818
Bipolar	0.769	0.730	0.749	0.822	0.704	0.759	0.804	0.739	0.770
BPD	0.787	0.785	0.786	**0.99**	0.788	0.793	0.751	0.809	0.779
Depression	0.726	0.861	0.788	0.739	0.838	0.785	**0.773**	0.793	0.783
None	0.887	0.877	0.882	**0.937**	0.859	0.896	0.900	0.881	0.890
OCD	0.899	0.873	0.885	0.900	0.891	0.896	**0.934**	0.871	0.901
PTSD	0.858	0.835	0.846	**0.912**	0.815	0.861	0.864	0.835	0.849
	MIRoBERTa			**Ensemble-BERT**			**Ensemble-RoBERTa**		
	P	R	F1	P	R	F1	P	R	F1
ADHD	0.899	0.898	0.898	0.896	0.870	0.883	**0.920**	0.884	**0.902**
Anxiety	0.845	0.837	**0.841**	0.824	0.843	0.833	0.825	0.847	0.836
Bipolar	0.779	**0.758**	0.768	**0.855**	0.707	0.774	0.820	0.737	**0.776**
BPD	0.872	0.712	0.784	0.772	**0.820**	**0.795**	0.850	0.745	0.794
Depression	0.721	0.878	0.792	0.757	0.842	0.797	0.738	**0.882**	**0.804**
None	0.902	**0.910**	0.906	0.894	0.902	0.898	0.905	**0.910**	**0.907**
OCD	0.899	0.891	0.895	0.915	**0.906**	**0.910**	0.904	**0.906**	0.905
PTSD	0.875	0.855	**0.865**	0.896	0.813	0.852	0.868	**0.860**	0.864

precision (0.837), recall (0.835), and F1-score (0.835) values at a learning rate of 1e-5. This performance surpasses that of MentalBERT [33], which has the next-best results with an F1-score of 0.832. The BioBERT model [19] follows closely with an F1-score of 0.821, whereas MedBERT [34] and ClinicalBERT [35] display lower performance with F1-scores of 0.804 and 0.799, respectively, at their best learning rates. These results imply that the domain-adaptive pretraining strategy for MIBERT captures the subtleties of mental health-related text, improving classification performance compared to other domain-specific BERT variants. The superior performance of MIBERT across all metrics emphasizes its potential as a robust tool for mental illness classification tasks.

B. ERROR DISCUSSION

Classifying mental illness, where various conditions, such as ADHD, depression, bipolar disorder, anxiety, BPD, obsessive-compulsive disorder (OCD), and PTSD are interconnected, achieving high accuracy with a multiclass classifier is challenging owing to overlapping symptoms and shared underlying factors [36]. For instance, bipolar disorder often co-occurs with other mental health diagnoses [37].

Fig. 3 compares the confusion matrices for the models, and Table 6 presents their classwise performance. The ensemble-RoBERTa model better identifies ADHD, with the highest precision (0.920) and F1-score (0.902) for this class. For anxiety classification, Longformer achieves the highest precision (0.860), while both BigBIRD and MIBERT share the highest recall (0.855), and MIRoBERTa achieves the highest F1-score (0.841). The bipolar class performance fluctuates across models, with ensemble-BERT achieving the highest precision (0.855), MIRoBERTa reaching the highest recall (0.758), and ensemble-RoBERTa obtaining the highest F1-score (0.776). For BPD classification, BigBIRD shows the highest precision (0.990), while ensemble-BERT achieves both the highest recall (0.820) and F1-score (0.795). Depression classification is effectively addressed by different models, with MIBERT showing the highest precision (0.773), ensemble-RoBERTa achieving the highest recall (0.882) and F1-score (0.804).

All models identify the ''none'' class well, with BigBIRD showing the highest precision (0.937), while both MIRoBERTa and ensemble-RoBERTa share the highest recall (0.910), and ensemble-RoBERTa achieves the highest F1-score (0.907). For OCD classification, MIBERT demonstrates the highest precision (0.934), while ensemble-BERT and ensemble-RoBERTa tie for the highest recall (0.906), and ensemble-BERT achieves the highest F1-score (0.910). In PTSD classification, BigBIRD shows the highest precision (0.912), ensemble-RoBERTa achieves the highest recall (0.860), and MIRoBERTa obtains the highest F1-score (0.865).

These results highlight the diverse strengths of each model across mental health conditions. The ensemble models, particularly ensemble-RoBERTa and ensemble-BERT, display strong overall performance across several classes. Domain-adaptive models like MIRoBERTa excel in certain metrics, while long-sequence transformers like BigBIRD demonstrate particular strengths in precision for several classes including BPD, None, and PTSD. This comparison highlights the complementary strengths of these advanced models in distinguishing between mental health conditions, even when conditions are interconnected.

C. WORD ANALYSIS

We analyzed word importance using Shapley additive explanations (SHAP) [38], [39] to interpret the MIRoBERTa transformer-based text classification model, quantifying the contribution of individual words to model predictions by assigning values reflecting the influence of each word across all input feature subsets. We analyzed the top 100 correctly predicted scores to identify which words significantly influenced model decisions. We removed explicit terms related to the label for each data point to avoid bias.

This approach uncovered subtler linguistic patterns and content that the model found influential in its classification process. The analysis revealed distinct sets of highly influential words for each mental health condition, as indicated in Fig. 4. Some words were split into subwords due to the transformer tokenizer. The analysis focused on complete words or meaningful subwords related to mental health conditions. The significant findings for each category are listed below:

Depression: The words ''defeat,'' ''despair,'' and ''lame'' had strong positive SHAP values, indicating their association with feelings of hopelessness, failure, and negative self-perception. The word ''comfort'' might reflect a desire for solace, whereas ''awake'' could relate to sleep disturbances common in depression.

Bipolar: The terms ''manic'' and ''invincible'' were highly influential, reflecting the cyclical nature of bipolar disorder, including manic episodes characterized by elevated mood and feelings of invincibility. ''Mixed'' likely refers to mixed episodes, whereas ''lithium'' points to a common medication used in treatment.

ADHD: The words ''clumsy,'' ''att'' (likely ''attention''), and ''forget'' were particularly influential, corresponding to various symptoms, such as poor motor coordination, attention deficits, and forgetfulness. ''Executive'' likely refers to executive functioning difficulties associated with ADHD.

BPD: Highly influential words included ''borderline,'' ''splitting,'' and ''abandonment,'' aligning with the core features of BPD, such as black-and-white thinking and intense fear of abandonment. ''Paranoid'' and ''manipulation'' reflect additional aspects of the disorder.

Anxiety: The analysis highlighted the words ''nervous,'' ''anxious,'' and ''worried,'' all directly related to the heightened state of worry characterizing anxiety disorders. ''Ochond'' (likely ''hypochondria'') suggests health-related anxiety, whereas ''pit'' might refer to physical symptoms, such as a ''pit in the stomach.''

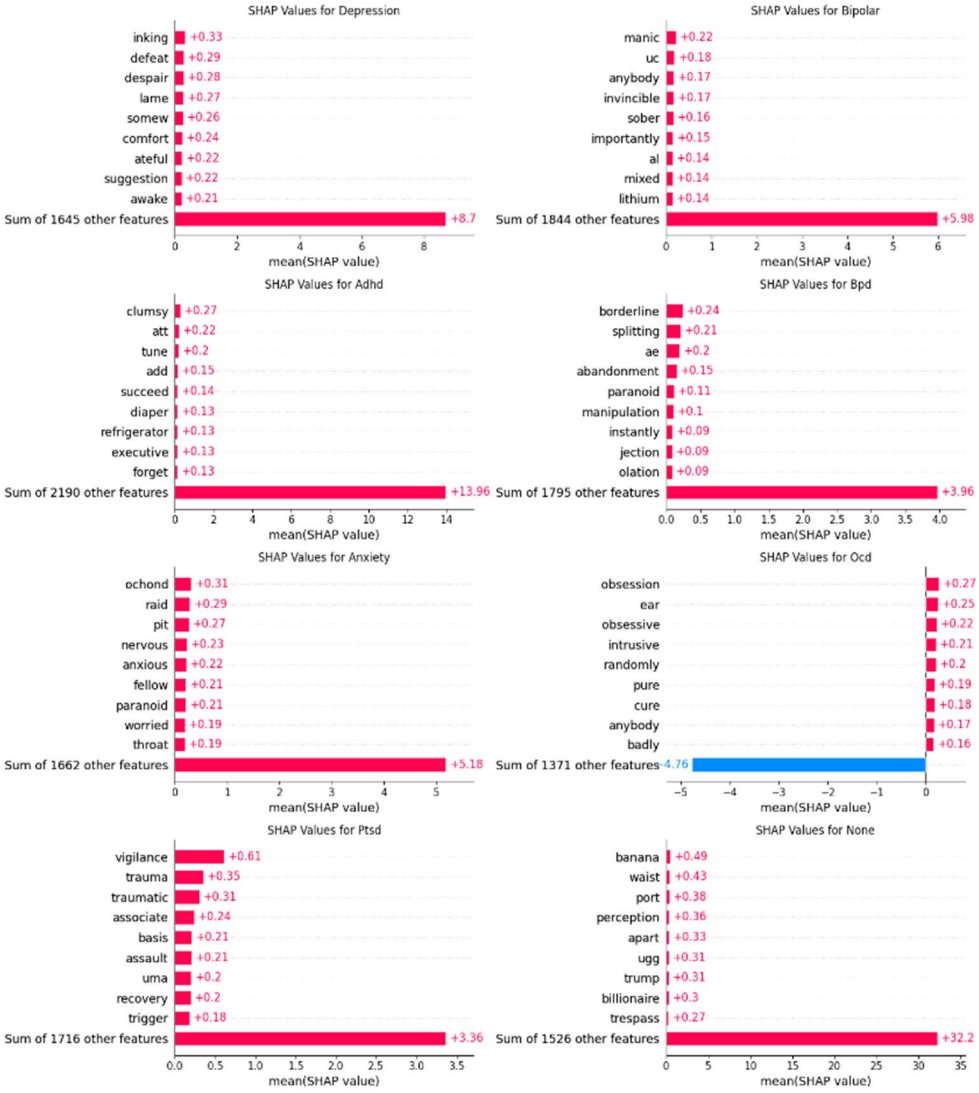

FIGURE 4. Shapley additive explanation (SHAP) values for mental health labels.

OCD: The words ''obsession,'' ''obsessive,'' and ''intrusive'' were highly influential, underscoring the repetitive, intrusive thoughts central to OCD. ''Pure'' might refer to ''Pure-O,'' a form of OCD focused on obsessions.

PTSD: The words ''vigilance,'' ''trauma,'' and ''assault'' were most influential, reflecting the hypervigilance, traumatic experiences, and potential causes of PTSD. ''Trigger'' relates to stimuli that can provoke PTSD symptoms.

None (unrelated terms): The ''none'' category included the words ''banana'' and ''billionaire,'' which had strong SHAP values but did not correspond to any mental health conditions, indicating their role in distinguishing unrelated content from mental health labels.

This analysis provides valuable insight into the linguistic features that the MIRoBERTa model associates with mental health conditions, clarifying its classification process.

D. MENTAL ILLNESS CLASSIFICATION APPLICATION

We developed a simple web application using the Streamlit framework to make the mental illness classification model accessible for public testing and feedback. Streamlit is a powerful Python library designed to build interactive, data-driven web apps rapidly. We deployed this application on Hugging Face Spaces using the ensemble method comprising ''roberta-base-finetuned-mental-health'' and ''mavinsao/mi-roberta-base-finetuned-mental-health''. The Hugging Face Spaces platform facilitates sharing and demonstrating machine learning models. This interface allows anyone to input text and receive real-time predictions, including the associated mental health label and confidence score. This simplified setup facilitates broader experimentation with the proposed model and offers valuable insight. Fig. 5 presents the web application highlighting text extracted from the online community on the r/depression subreddit. The application accurately identified

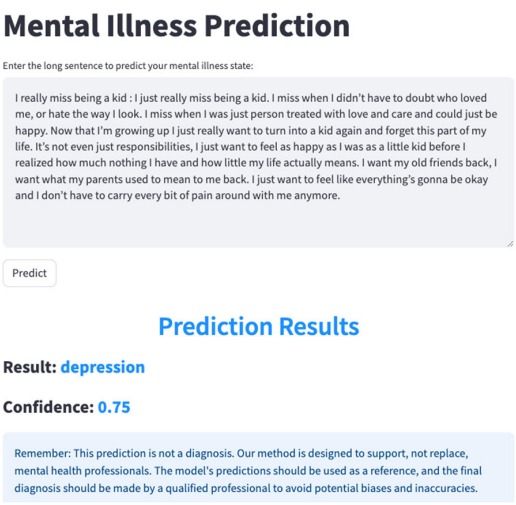

FIGURE 5. Streamlit web application demonstration.

the label as depression despite the lack of the explicit term ''depression'' in the text. The space application is available at https://huggingface.co/spaces/mavinsao/ mental-illness-classification.

VII. CONCLUSION

This study introduces MIRoBERTa-base and MIBERT-base-uncased, two new language models designed for mental health topics. Overall, MIRoBERTa outperforms other models on this task. The pretrained BERT model (MIBERT) surpassed existing domain-specific pretrained models, including ClinicalBERT, MedBERT, MentalBERT, and BioBERT. Our ensemble learning strategy, combining MIRoBERTa with RoBERTa and MIBERT with BERT, sets a new standard in mental health text classification, achieving superior results compared to individual models. We created a comprehensive dataset with 54 161 submissions related to mental health to support future studies. We also applied SHapley Additive exPlanations (SHAP) to interpret model decisions, providing critical insights into the inner workings of these complex models, which is essential in sensitive domains like mental health. Finally, we developed a simple web application on the Hugging Face Spaces platform where testers can obtain mental health predictions based on textual input.

The current work acknowledges the limitations of multilabel datasets and advises that future research should assess more tailored approaches to multilabel classification. Expanding the model scope by incorporating training on multilingual datasets is recommended, which could substantially improve its applicability across linguistic groups.

- **MIRoBERTa-base**: https://huggingface.co/mavinsao/mi-roberta-base
- **Mental Health Dataset**: https://drive.google.com/drive/folders/1xX5cTx21qjhFAF4giHg_TJIiPs4mIZWQ?usp=sharing
- **Project Repository**: https://github.com/MavinSao/mentalhealth-text-classification

This study is an initial step toward developing artificial intelligence-driven mental health support systems for text-based communication. These systems can enhance the accessibility and effectiveness of mental health therapy in online environments by identifying prospective indicators of mental health concerns in user interactions and composing appropriate, personalized responses.

A. LIMITATIONS

High accuracy in multiclass mental illness classification is challenging due to the complex interconnections between mental illnesses. Employing Reddit subreddits as a data source introduces further complexity because the labels might not consistently align with standardized clinical mental health diagnoses. Although we required submissions with at least ten upvotes, user posts may be irrelevant to the subreddit topic. To improve the dataset in the future, we plan to validate labels via domain-expert human validation, if feasible.

B. ETHICAL CONSIDERATIONS

Privacy and Confidentiality: Given the sensitive nature of mental health discussions, anonymizing the data to protect the identities of the individuals who share their experiences on Reddit is crucial. This protection involves removing personally identifiable information and ensuring the data cannot be traced to the original users. All personal details, such as the username, link address, date, time, and user identification, are removed or anonymized, eliminating the risk of personal information exposure.

Informed Consent: When conducting research using data from Reddit, obtaining direct consent from each user who posted content is often challenging. Instead, we followed Reddit's rules and ethical guidelines for using publicly available data.

Role of AI in Diagnosis: This method is designed to assist mental health professionals, such as psychologists and psychiatrists, rather than to replace them. Using the proposed model to calculate mental illness labels directly can introduce bias, potentially leading to inaccurate diagnoses. Therefore, the predictions made using the proposed model should only be used as a reference, with qualified professionals carefully determining the final diagnosis.

REFERENCES

[1] H. Legido-Quigley, N. Asgari, Y. Y. Teo, G. M. Leung, H. Oshitani, K. Fukuda, A. R. Cook, L. Y. Hsu, K. Shibuya, and D. Heymann, ''Are high-performing health systems resilient against the COVID-19 epidemic?'' *Lancet*, vol. 395, no. 10227, pp. 848–850, Mar. 2020, doi: 10.1016/s0140-6736(20)30551-1.

[2] M. Nicola, Z. Alsafi, C. Sohrabi, A. Kerwan, A. Al-Jabir, C. Iosifidis, M. Agha, and R. Agha, ''The socio-economic implications of the coronavirus pandemic (COVID-19): A review,'' *Int. J. Surg.*, vol. 78, pp. 185–193, Jun. 2020, doi: 10.1016/j.ijsu.2020.04.018.

[3] E. A. Holmes, ''Multidisciplinary research priorities for the COVID-19 pandemic: A call for action for mental health science,'' *Lancet Psychiatry*, vol. 7, no. 6, pp. 547–560, Jun. 2020, doi: 10.1016/s2215-0366(20)30168-1.

[4] R. Bauer, M. Bauer, H. Spiessl, and T. Kagerbauer, ''Cyber-support: An analysis of online self-help forums (online self-help forums in bipolar disorder),'' *Nordic J. Psychiatry*, vol. 67, no. 3, pp. 185–190, Jul. 2012, doi: 10.3109/08039488.2012.700734.

[5] S. J. Sowles, M. J. Krauss, L. Gebremedhn, and P. A. Cavazos-Rehg, "'I feel like I've hit the bottom and have no idea what to do': Supportive social networking on Reddit for individuals with a desire to quit cannabis use," *Substance Abuse*, vol. 38, no. 4, pp. 477–482, Oct. 2017, doi: 10.1080/08897077.2017.1354956.

[6] I. Herrera-Peco, I. Fernández-Quijano, and C. Ruiz-Núñez, "The role of social media as a resource for mental health care," *Eur. J. Invest. Health, Psychol. Educ.*, vol. 13, no. 6, pp. 1026–1028, Jun. 2023, doi: 10.3390/ejihpe13060078.

[7] PullPush Reddit API. (2024). *Pullpush Reddit API Documentation*. Accessed: Mar. 28, 2024. [Online]. Available: https://pullpush.io/

[8] S. Ji, C. P. Yu, S.-F. Fung, S. Pan, and G. Long, "Supervised learning for suicidal ideation detection in online user content," *Complexity*, vol. 2018, no. 1, pp. 1–10, Sep. 2018, doi: 10.1155/2018/6157249.

[9] A. H. Orabi, P. Buddhitha, M. H. Orabi, and D. Inkpen, "Deep learning for depression detection of Twitter users," in *Proc. 5th Workshop Comput. Linguistics Clin. Psychol., From Keyboard Clinic*, 2018, pp. 88–97, doi: 10.18653/v1/w18-0609.

[10] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Detection of depression-related posts in Reddit social media forum," *IEEE Access*, vol. 7, pp. 44883–44893, 2019, doi: 10.1109/ACCESS.2019.2909180.

[11] P. Arora and P. Arora, "Mining Twitter data for depression detection," in *Proc. Int. Conf. Signal Process. Commun. (ICSC)*, Noida, India, Mar. 2019, pp. 186–189, doi: 10.1109/ICSC45622.2019.8938353.

[12] V. Borba de Souza, J. Campos Nobre, and K. Becker, "DAC stacking: A deep learning ensemble to classify anxiety, depression, and their comorbidity from Reddit texts," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 7, pp. 3303–3311, Jul. 2022, doi: 10.1109/JBHI.2022.3151589.

[13] T. H. H. Aldhyani, S. N. Alsubari, A. S. Alshebami, H. Alkahtani, and Z. A. T. Ahmed, "Detecting and analyzing suicidal ideation on social media using deep learning and machine learning models," *Int. J. Environ. Res. Public Health*, vol. 19, no. 19, p. 12635, Oct. 2022, doi: 10.3390/ijerph191912635.

[14] D. M. Low, L. Rumker, T. Talkar, J. Torous, G. Cecchi, and S. S. Ghosh, "Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on Reddit during COVID-19: Observational study," *J. Med. Internet Res.*, vol. 22, no. 10, Oct. 2020, Art. no. e22635, doi: 10.2196/22635.

[15] C. Yu, F. Wang, Z. Shao, T. Qian, Z. Zhang, W. Wei, and Y. Xu, "GinAR: An end-to-end multivariate time series forecasting model suitable for variable missing," 2024, *arXiv:2405.11333*.

[16] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A time series is worth 64 words: Long-term forecasting with transformers," 2022, *arXiv:2211.14730*.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North American Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, Jun. 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.

[18] A. Murarka, B. Radhakrishnan, and S. Ravichandran, "Classification of mental illnesses on social media using RoBERTa," in *Proc. 7th Int. Workshop Lang. Understand. Healthcare*, Apr. 2021, pp. 59–68. [Online]. Available: https://aclanthology.org/2021.louhi-1.7/

[19] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Sep. 2019, doi: 10.1093/bioinformatics/btz682.

[20] M. Rizwan, M. F. Mushtaq, U. Akram, A. Mehmood, I. Ashraf, and B. Sahelices, "Depression classification from tweets using small deep transfer learning language models," *IEEE Access*, vol. 10, pp. 129176–129189, 2022, doi: 10.1109/ACCESS.2022.3223049.

[21] R. Poświata and M. Perełkiewicz, "OPI@LT-EDI-ACL2022: Detecting signs of depression from social media text using RoBERTa pre-trained language models," in *Proc. 2nd Workshop Lang. Technol. Equality, Diversity Inclusion*, Jan. 2022, pp. 276–282, doi: 10.18653/v1/2022.ltedi-1.40.

[22] Y. Qiao, C. Xiong, Z. Liu, and Z. Liu, "Understanding the behaviors of BERT in ranking," 2019, *arXiv:1904.07531*.

[23] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manage.*, vol. 24, no. 5, pp. 513–523, Jan. 1988, doi: 10.1016/0306-4573(88)90021-0.

[24] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Oct. 2014, pp. 1532–1543, doi: 10.3115/v1/D14-1162.

[25] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[26] M. Zaheer, G. Guruganesh, A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, and A. Ahmed, "Big bird: Transformers for longer sequences," 2020, *arXiv:2007.14062*.

[27] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," 2020, *arXiv:2004.05150*.

[28] Zenodo. (2024). *Reddit Mental Health Dataset*. Accessed: Mar. 28, 2024. [Online]. Available: https://zenodo.org/records/3941387

[29] U. Naftaly, N. Intrator, and D. Horn, "Optimal ensemble averaging of neural networks," *Network: Comput. Neural Syst.*, vol. 8, no. 3, pp. 283–296, Jan. 1997, doi: 10.1088/0954-898x_8_3_004.

[30] A. Paszke, "PyTorch: An imperative style, high-performance deep learning library," 2019, *arXiv:1912.01703*.

[31] T. Wolf, "HuggingFace's transformers: State-of-the-art natural language processing," 2019, *arXiv:1910.03771*.

[32] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014. [Online]. Available: https://www.cs.toronto.edu/

[33] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria, "MentalBERT: Publicly available pretrained language models for mental healthcare," 2021, *arXiv:2110.15621*.

[34] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, and D. Zhi, "Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction," 2020, *arXiv:2005.12833*.

[35] K. Huang, J. Altosaar, and R. Ranganath, "ClinicalBERT: Modeling clinical notes and predicting hospital readmission," 2019, *arXiv:1904.05342*.

[36] J. L. Doherty and M. J. Owen, "Genomic insights into the overlap between psychiatric disorders: Implications for research and clinical practice," *Genome Med.*, vol. 6, no. 4, Apr. 2014, Art. no. 29, doi: 10.1186/gm546.

[37] T. Hvilivitzky. (Aug. 4, 2023). *7 Conditions That Can Go Hand in Hand With Bipolar Disorder*. bpHope.com. Accessed: Apr. 3, 2024. [Online]. Available: https://www.bphope.com/bipolar-buzz/6-conditions-that-can-go-hand-in-hand-with-bipolar-disorder

[38] *SHapley Additive ExPlanations*. Accessed: Aug. 29, 2024. [Online]. Available: https://shap.readthedocs.io/en/latest/index.html

[39] S. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.* New York, NY, USA: Curran Associates, Jan. 2017, pp. 4765–4774.

MAVIN SAO received the B.A. degree from the Royal University of Phnom Penh, Cambodia, in 2020. He is currently pursuing the master's degree in data science with Chonnam National University.

From 2020 to 2022, he worked as an IT Instructor at the Global NGO Foundation for Korea Software Global Aid in Phnom Penh. His research interests include LLM, data mining, and natural language processing.

HOI-JEONG LIM received the combined M.S. and Ph.D. degree from Columbia University, New York, NY, USA, in 2020.

From 1998 to 2000, she worked as a Biostatistician at the Neurological Institute of New York, Columbia University Medical Center. From 2001 to 2004, she worked as a Postdoctoral Research Scientist at the School of Medicine, Seoul National University. From 2005 to 2021, she was a Professor at the School of Dentistry, Chonnam National University. Since 2022, she has been working as a Professor with the Graduate School of Data Science. She is concurrently working as the Director of the Public Data Analytics Center, Chonnam National University. Most recently, her work has focused on large language model applications.